Deep Learning (CSE641/ECE555)
Quiz-2 (15 Marks) (Duration 60 min) *

Name ......
Roll No. ......

Question 1-12 [1 Marks], Question 13 [3 Marks]

1. In the Transformer model, what is the purpose of the positional encoding?

   (a) To add noise to the input embeddings for regularization.
   (b) To provide information about the position of tokens in the sequence.
   (c) To reduce the dimensionality of the input embeddings.
   (d) To initialize the weights of the self-attention layers.

   B

2. What is the time complexity of the self-attention mechanism in a Transformer for a sequence of length n? (a) $O(n)$ (b) $O(n \log n)$ (c) $O(n^2)$ (d) $O(1)$ C

3. In an LSTM, the cell state c_t at time t is updated using the forget gate f_t and the input gate i_t. If c_{t-1} is the previous cell state and $\tilde{c}_t$ is the candidate cell state, what is the update formula for c_t? (a) $c_t = f_t \cdot c_{t-1} + i_t \cdot \tilde{c}_t$ (b) $c_t = f_t \cdot c_{t-1} \cdot i_t \cdot \tilde{c}_t$ (c) $c_t = f_t + i_t \cdot \tilde{c}_t$ (d) $c_t = f_t \cdot \tilde{c}_t + i_t \cdot c_{t-1}$ A

4. In self-attention, given scaled dot-product scores [2, 0, -1], what are the attention weights after softmax?

   (a) (0.88, 0.11, 0.01)
   (b) (0.5, 0.3, 0.2)
   (c) (0.7, 0.2, 0.1)
   (d) (0.9, 0.05, 0.05)

   A

5. Which of the following statements is INCORRECT?

   (a) Recurrent neural networks can handle a sequence of arbitrary length, while feedforward neural networks can not.
   (b) Training recurrent neural networks is hard because of vanishing and exploding gradient problems.
   (c) Gated recurrent units (GRUs) have fewer parameters than LSTMs.
   (d) Gradient clipping is an effective way of solving vanishing gradient problem.

   D

6. In transformer decoder architectures, the causality constraint in attention is often implemented by applying a mask to the attention logits. What is the computational advantage of this approach compared to a pure sequential autoregressive implementation using a `for` loop?

   (a) It enables all positions to be processed in parallel rather than sequentially
   (b) It completely eliminates the need for attention mechanisms
   (c) It reduces the sequence length by trimming unnecessary tokens
   (d) It reduces the number of parameters in the model

   C. Using matrix operations with masking allows the model to compute attention scores for all tokens simultaneously rather than generating each token one by one in a sequential loop. This parallelization significantly speeds up both training and inference compared to truly autoregressive implementations like traditional RNNs/LSTMs.

7. In a decoder-only transformer model that employs causal attention over a sequence of length L, what are the maximum dimensions that the attention mask matrix can have? (a) vocab_size x L (b) L × L (c) batch_size × L × L (d) L × vocab_size B. The causal attention mask matrix has dimensions L × L, where L is the context length. This creates a lower triangular matrix where each token can attend to itself and all previous tokens but not to future tokens.

8. In deep RNNs, which mathematical property of the sigmoid and tanh activation functions primarily contributes to the vanishing gradient problem? □

   (a) Their output range is bounded between 0 and 1 (or -1 and 1)

   (b) Their derivatives have maximum values less than 1

   (c) Their derivatives approach zero for very large or very small inputs

   (d) All of the above

   C.

9. Which of these is not a good criterion for a good positional encoding algorithm? □

   (a) It should output a common encoding for each time-step.

   (b) Distance between any two time-steps should be consistent for all sentence lengths.

   (c) It must be deterministic.

   (d) The algorithm should be able to generalize to longer sentences.

   A

10. Which of the following is the correct formula for gradient clipping? □

    (a) $\hat{g} \leftarrow \hat{g}^2 + \text{threshold}$ if $\|\hat{g}\| \geq$ threshold

    (b) $\hat{g} \leftarrow c$ if $\|\hat{g}\| \geq$ threshold where c is a hyper-parameter

    (c) $\hat{g} \leftarrow \frac{\text{threshold}}{\|\hat{g}\|}\hat{g}$ if $\|\hat{g}\| \geq$ threshold

    (d) $\hat{g} \leftarrow \text{ReLU}(\hat{g})$ to remove negative gradients

    C

11. How many gates does a GRU (Gated Recurrent Unit) cell have? □

    (a) 0; there is no gating

    (b) 1; forget gate

    (c) 2; reset and update gates

    (d) 3; reset, forget, and update gates

    C

12. Which mathematical operation is used to implement the gating mechanisms in a GRU? (a) Matrix addition (b) Matrix multiplication (Dot product) (c) Element-wise multiplication (Hadamard product) (d) Convolution (Discrete convolution) C □

13. Consider a simple Recurrent Neural Network (RNN) for token classification with the following definitions:

    - **Input at time step** t: $x_t \in \mathbb{R}^{n_x}$.
    - **Hidden state**: $s_t \in \mathbb{R}^{n_h}$.
    - **Output**: $y_t \in \mathbb{R}^{n_y}$.
    - **Weight matrices**:
      - $W \in \mathbb{R}^{n_h \times n_x}$ (input-to-hidden weights),
      - $U \in \mathbb{R}^{n_h \times n_h}$ (hidden-to-hidden weights),
      - $V \in \mathbb{R}^{n_y \times n_h}$ (hidden-to-output weights).

    The forward pass equations are:

    $$s_t = tanh(Us_{t-1} + Wx_t), \tag{1}$$

    $$y_t = Vs_t, \tag{2}$$

    where $\phi(s)$ is an activation function, typically tanh.

    The loss function is defined as:

    $$L = \sum_t L_t = \sum_t \ell(y_t, \hat{y}_t). \tag{3}$$

    Using Backpropagation Through Time (BPTT), derive the gradient of the loss function with respect to V.

Answer: The loss function is:

$$L = \sum_t L_t = \sum_t \ell(y_t, \hat{y}_t) \tag{4}$$

The gradient with respect to V is computed as:

$$\frac{\partial L}{\partial V} = \sum_t \frac{\partial L_t}{\partial y_t} s_t^T \tag{5}$$